Chapter 11
Specification Error Analysis

The specification of a linear regression model consists of a formulation of the regression relationships and of statements or assumptions concerning the explanatory variables and disturbances. If any of these is violated, e.g., incorrect functional form, the improper introduction of disturbance term in the model, etc., then specification error occurs. In a narrower sense, the specification error refers to explanatory variables.

The complete regression analysis depends on the explanatory variables present in the model. It is understood in the regression analysis that only correct and important explanatory variables appear in the model. In practice, after ensuring the correct functional form of the model, the analyst usually has a pool of explanatory variables which possibly influence the process or experiment. Generally, all such candidate variables are not used in the regression modeling, but a subset of explanatory variables is chosen from this pool.

When choosing a subset of explanatory variables, there are two possible options:

1. To make the model as realistic as possible, the analyst may include as many as possible explanatory variables.
2. To make the model as simple as possible, one may include fewer explanatory variables.

In such selections, there can be two types of incorrect model specifications.

1. Omission/exclusion of relevant variables.
2. Inclusion of irrelevant variables.

Now we discuss the statistical consequences arising from both situations.

1. Exclusion of relevant variables:

To keep the model simple, the analyst may delete some of the explanatory variables which may be of importance from the point of view of theoretical considerations. There can be several reasons behind such decisions, e.g., it may be hard to quantify the variables like taste, intelligence, etc. Sometimes it may be difficult to make correct observations on the variables like income etc.

Let there be k candidate explanatory variables out of which suppose r variables are included and $(k-r)$ variables are to be deleted from the model. So partition the X and β as

$$\underset{n\times k}{X} = \left(\underset{n\times r}{X_1} \quad \underset{n\times(k-r)}{X_2} \right) \quad \text{and} \quad \beta = \left(\underset{r\times 1}{\beta_1} \quad \underset{(k-r)\times 1)}{\beta_2} \right).$$

The model $y = X\beta + \varepsilon,\ E(\varepsilon) = 0,\ V(\varepsilon) = \sigma^2 I$ can be expressed as

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

which is called a **full model** or a **true model.**

After retaining r explanatory variables and dropping $(k-r)$ explanatory variables the in the model, the new model is

$$y = X_1\beta_1 + \delta$$

which is called a **misspecified mode**l or **false model.**

Applying OLS to the false model, the OLSE of β_1 is

$$b_{1F} = (X_1'X_1)^{-1}X_1'y.$$

The estimation error is obtained as follows:

$$\begin{aligned} b_{1F} &= (X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2 + \varepsilon) \\ &= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'\varepsilon \\ b_{1F} - \beta_1 &= \theta + (X_1'X_1)^{-1}X_1'\varepsilon \end{aligned}$$

where $\theta = (X_1'X_1)^{-1}X_1'X_2\beta_2$.

Thus

$$\begin{aligned} E(b_{1F} - \beta_1) &= \theta + (X_1'X_1)^{-1}E(\varepsilon) \\ &= \theta \end{aligned}$$

which is a linear function of β_2, i.e., the coefficients of excluded variables. So b_{1F} is biased, in general. The bias vanishes if $X_1'X_2 = 0,$ i.e., X_1 and X_2 are orthogonal or uncorrelated.

The mean squared error matrix of b_{1F} is

$$\begin{aligned} MSE(b_{1F}) &= E(b_{1F} - \beta_1)(b_{1F} - \beta_1)' \\ &= E\left[\theta\theta' + \theta\varepsilon' X_1(X_1'X_1)^{-1} + (X_1'X_1)^{-1}X_1'\varepsilon\theta' + (X_1'X_1)^{-1}X_1'\varepsilon\varepsilon' X_1(X_1'X_1)^{-1}\right] \\ &= \theta\theta' + 0 + 0 + \sigma^2(X_1'X_1)^{-1}X_1'IX_1(X_1'X_1)^{-1} \\ &= \theta\theta' + \sigma^2(X_1'X_1)^{-1}. \end{aligned}$$

So efficiency generally declines. Note that the second term is the conventional form of MSE.

The residual sum of squares is

$$\hat{\sigma}^2 = \frac{SS_{res}}{n-r} = \frac{e'e}{n-r}$$

where $e = y - X_1 b_{1F} = \bar{H}_1 y,$

$$\bar{H}_1 = I - X_1(X_1' X_1)^{-1} X_1'.$$

Thus

$$\begin{aligned}\bar{H}_1 y &= \bar{H}_1(X_1\beta_1 + X_2\beta_2 + \varepsilon)\\ &= 0 + \bar{H}_1(X_2\beta_2 + \varepsilon)\\ &= \bar{H}_1(X_2\beta_2 + \varepsilon).\end{aligned}$$

$$\begin{aligned}y'\bar{H}_1 y &= (X_1\beta_1 + X_2\beta_2 + \varepsilon)\bar{H}_1(X_2\beta_2 + \varepsilon)\\ &= (\beta_2' X_2 \bar{H}_1' \bar{H}_1 X_2 \beta_2 + \beta_2' X_2' \bar{H}_1 \varepsilon + \beta_2' X_2 \bar{H}_1' X_2 \beta_2 + \beta_1' X_1' \bar{H}_1 \varepsilon + \varepsilon' \bar{H}_1' X_2 \beta_2 + \varepsilon' \bar{H}_1 \varepsilon).\end{aligned}$$

$$\begin{aligned}E(s^2) &= \frac{1}{n-r}\left[E(\beta_2' X_2' \bar{H}_1 X_2 \beta_2) + 0 + 0 + E(\varepsilon' \bar{H}\varepsilon)\right]\\ &= \frac{1}{n-r}\left[\beta_2' X_2' \bar{H}_1 X_2 \beta_2) + (n-r)\sigma^2\right]\\ &= \sigma^2 + \frac{1}{n-r}\beta_2' X_2' \bar{H}_1 X_2 \beta_2.\end{aligned}$$

Thus s^2 is a biased estimator of σ^2 and s^2 provides an overestimate of σ^2. Note that even if $X_1' X_2 = 0$, then also s^2 gives an overestimate of σ^2. So the statistical inferences based on this will be faulty. The t-test and confidence region will be invalid in this case.

If the response is to be predicted at $x' = (x_1', x_2'),$ then using the full model, the predicted value is

$$\hat{y} = x'b = x'(X'X)^{-1}X'y$$

with

$$\begin{aligned}E(\hat{y}) &= x'\beta\\ Var(\hat{y}) &= \sigma^2\left[1 + x'(X'X)^{-1}x\right].\end{aligned}$$

When the subset model is used then the predictor is

$$\hat{y}_1 = x_1' b_{1F}$$

and then

$$
\begin{aligned}
E(\hat{y}_1) &= x_1'(X_1'X_1)^{-1}X_1'E(y) \\
&= x_1'(X_1'X_1)^{-1}X_1'E(X_1\beta_1 + X_2\beta_2 + \varepsilon) \\
&= x_1'(X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2) \\
&= x_1'\beta_1 + x_1'(X_1'X_1)^{-1}X_1'X_2\beta_2 \\
&= x_1'\beta_1 + x_i'\theta.
\end{aligned}
$$

Thus $\hat{y}_1$ is a biased predictor of y. It is unbiased when $X_1'X_2 = 0.$ The MSE of predictor is

$$MSE(\hat{y}_1) = \sigma^2\left[1 + x_1'(X_1'X_1)^{-1}x_1\right] + \left(x_1'\theta - x_2'\beta_2\right)^2.$$

Also

$$Var(\hat{y}) \geq MSE(\hat{y}_1)$$

provided $V(\hat{\beta}_2) - \beta_2\beta_2'$ is positive semidefinite.

2. Inclusion of irrelevant variables

Sometimes due to enthusiasm and to make the model more realistic, the analyst may include some explanatory variables that are not very relevant to the model. Such variables may contribute very little to the explanatory power of the model. This may tend to reduce the degrees of freedom $(n-k)$, and consequently, the validity of the inference drawn may be questionable. For example, the coefficient of determination will increase, indicating that the model is improving, which may not be true.

Let the true model be

$$y = X\beta + \varepsilon,\ E(\varepsilon) = 0, V(\varepsilon) = \sigma^2 I$$

which comprise k explanatory variable. Suppose now r additional explanatory variables are added to the model and the resulting model becomes

$$y = X\beta + Z\gamma + \delta$$

where Z is a $n \times r$ matrix of n observations on each of the r explanatory variables and γ is $r \times 1$ vector of regression coefficient associated with Z and δ is disturbance term. This model is termed as a **false model.**

Applying OLS to a false model, we get

$$\begin{pmatrix} b_F \\ c_F \end{pmatrix} = \begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z \end{pmatrix}^{-1} \begin{pmatrix} X'y \\ Z'y \end{pmatrix}$$

$$\begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z \end{pmatrix} \begin{pmatrix} b_F \\ c_F \end{pmatrix} = \begin{pmatrix} X'y \\ Z'y \end{pmatrix}$$

$$\Rightarrow X'Xb_F + X'Zc_F = X'y \qquad (1)$$

$$Z'Xb_F + Z'Zc_F = Z'y \qquad (2)$$

where b_F and C_F are the OLSEs of β and γ respectively.

Premultiply equation (2) by $X'Z(Z'Z)^{-1}$, we get

$$X'Z(Z'Z)^{-1}Z'Xb_F + X'Z(Z'Z)^{-1}Z'Zc_F = X'Z(Z'Z)^{-1}Z'y. \qquad (3)$$

Subtracting equation (1) from (3), we get

$$\left[X'X - X'Z(Z'Z)^{-1}Z'X\right]b_F = X'y - X'Z(Z'Z)^{-1}Z'y$$

$$X'\left[I - Z(Z'Z)^{-1}Z'\right]Xb_F = X'\left[I - Z(Z'Z)^{-1}Z'\right]y$$

$$\Rightarrow b_F = (X'\bar{H}_Z X)^{-1}X'\bar{H}_Z y$$

where $\bar{H}_Z = I - Z(Z'Z)^{-1}Z'.$

The estimation error of b_F is

$$\begin{aligned} b_F - \beta &= (X'\bar{H}_Z X)^{-1}X'\bar{H}_Z y - \beta \\ &= (X'\bar{H}_Z X)^{-1}X'\bar{H}_Z (X\beta + \varepsilon) - \beta \\ &= (X'\bar{H}_Z X)^{-1}X'\bar{H}_Z \varepsilon. \end{aligned}$$

Thus

$$E(b_F - \beta) = (X'\bar{H}_Z X)^{-1}X'\bar{H}_Z E(\varepsilon) = 0$$

so b_F is unbiased even when some irrelevant variables are added to the model.

The covariance matrix is

$$\begin{aligned} V(b_F) &= E\left(b_F - \beta\right)\left(b_F - \beta\right)^1 \\ &= E\left[\left(X'\bar{H}_Z X\right)^{-1} X'\bar{H}_Z \varepsilon\varepsilon'\bar{H}_Z X(X'\bar{H}_Z X)^{-1}\right] \\ &= \sigma^2\left(X'\bar{H}_Z X\right)^{-1} X'\bar{H}_Z I\bar{H}_Z X\left(X'\bar{H}_Z X\right)^{-1} \\ &= \sigma^2\left(X'\bar{H}_Z X\right)^{-1}. \end{aligned}$$

If OLS is applied to true model, then

$$b_T = (X'X)^{-1}X'y$$

with $E(b_T) = \beta$

$$V(b_T) = \sigma^2 (X'X)^{-1}.$$

To compare b_F and b_T, we use the following result.

Result: If A and B are two positive definite matrices then $A - B$ is at least positive semi-definite if $B^{-1} - A^{-1}$ is also at least positive semi-definite.

Let

$$\begin{aligned}
A &= (X'\bar{H}_Z X)^{-1} \\
B &= (X'X)^{-1} \\
B^{-1} - A^{-1} &= X'X - X'\bar{H}_Z X \\
&= X'X - X'X + X'Z(Z'Z)^{-1}Z'X \\
&= X'Z(Z'Z)^{-1}Z'X
\end{aligned}$$

which is at least a positive semi-definite matrix. This implies that the efficiency declines unless $X'Z = 0.$ If $X'Z = 0,$ i.e., X and Z are orthogonal, then both are equally efficient.

The residual sum of squares under the false model is

$$SS_{res} = e_F' e_F$$

where

$$\begin{aligned}
e_F &= y - Xb_F - ZC_F \\
b_F &= (X'\bar{H}_Z X)^{-1} X'\bar{H}_Z y \\
c_F &= (Z'Z)^{-1}Z'y - (Z'Z)^{-1}Z'Xb_F \\
&= (Z'Z)^{-1}Z'(y - Xb_F) \\
&= (Z'Z)^{-1}Z'\left[I - X(X'\bar{H}_Z X)^{-1}X'\bar{H}_z\right]y \\
&= (Z'Z)^{-1}Z'\bar{H}_{XZ}y \\
\bar{H}_Z &= I - Z(Z'Z)^{-1}Z' \\
\bar{H}_{Zx} &= I - X(X'\bar{H}_Z X)^{-1}X'\bar{H}_Z \\
\bar{H}_{ZX}^2 &= \bar{H}_{ZX}: \text{ idempotent.}
\end{aligned}$$

So

$$\begin{aligned} e_F &= y - X(X'\bar{H}_Z X)^{-1} X'\bar{H}_Z y - Z(Z'Z)^{-1} Z'\bar{H}_{ZX} y \\ &= \left[I - X(X'\bar{H}_Z X)^{-1} X'\bar{H}_Z - Z(Z'Z)^{-1} Z' H_{ZX} \right] y \\ &= \left[\bar{H}_{ZX} - (I - \bar{H}_Z)\bar{H}_{ZX} \right] y \\ &= \bar{H}_Z \bar{H}_{ZX} y \\ &= \bar{H}^*_{ZX} y \;\; \text{where } \; \bar{H}^*_{ZX} = \bar{H}_Z \bar{H}_{ZX}. \end{aligned}$$

Thus

$$\begin{aligned} SS_{res} &= e_F' e_F \\ &= y'\bar{H}_Z \bar{H}_{ZX} \bar{H}_{ZX} \bar{H}_Z y \\ &= y'\bar{H}_Z \bar{H}_{ZX} y \\ &= y'\bar{H}^*_{ZX} y \end{aligned}$$

$$\begin{aligned} E(SS_{res}) &= \sigma^2 \, tr(\bar{H}^*_{ZX}) \\ &= \sigma^2 (n-k-r) \end{aligned}$$

$$E\left(\frac{SS_{res}}{n-k-r} \right) = \sigma^2.$$

So $\dfrac{SS_{res}}{n-k-r}$ is an unbiased estimator of σ^2.

A comparison of exclusion and inclusion of variables is as follows:

	Exclusion type	**Inclusion type**
Estimation of coefficients	Biased	Unbiased
Efficiency	Generally declines	Declines
Estimation of the disturbance term	Over-estimate	Unbiased
Conventional test of hypothesis and confidence region	Invalid and faulty inferences	Valid though erroneous